DOCUMENT RESUME

ED 418 130 TM 028 224

AUTHOR Li, Yuan H.; Griffith, William D.; Tam, Hak P.
TITLE Equating Multiple Tests via an IRT Linking Design: Utilizing a Single Set of Anchor Items with Fixed Common Item Parameters during the Calibration Process.
PUB DATE 1997-06-00
NOTE 47p.; Paper presented at the Annual Meeting of the Psychometric Society (Knoxville, TN, June 26-29, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; Difficulty Level; *Equated Scores; Error of Measurement; *Estimation (Mathematics); *Item Response Theory; Simulation; Tables (Data); Test Format
IDENTIFIERS Anchor Tests; Calibration; Item Parameters; *Linkage

ABSTRACT

This study explores the relative merits of a potentially useful item response theory (IRT) linking design: using a single set of anchor items with fixed common item parameters (FCIP) during the calibration process. An empirical study was conducted to investigate the appropriateness of this linking design using 6 groups of students taking 6 forms of a pilot test, for an accumulated sample size of 8,357 students. A parameter recovery study was performed to examine the robustness of FCIP under the situation of large standard errors in the item difficulty and guessing parameters. Comparison of these results to those produced by the characteristics curve method (CCM) was pursued. Based on the empirical portion of this study, ability estimates calibrated from this linking design are very consistent, except for students with extreme (especially low) ability under the CCM equating method. Item parameter estimates calibrated from this linking design are also very consistent, except for guessing parameters under the CCM equating method. Based on the results from the simulation portion of the study, this linking result can produce very precise and stable parameter estimates. (Contains 7 tables, 19 figures, and 49 references.) (Author/SLD)

Equating Multiple Tests via an IRT Linking Design: Utilizing a Single Set of Anchor Items with Fixed Common Item Parameters during the Calibration Process

*Yuan H. Li
Prince George's County Public School, Maryland
University of Maryland

William D. Griffith
Prince George's County Public School, Maryland

Hak P. Tam
National Taiwan Normal University

Paper presented at the 1997 Psychometric Society Meeting
June 26-29, Knoxville, TN

* Graduate Program at the University of Maryland and employed by the Prince George's County Public Schools, Maryland

Equating Multiple Tests via an IRT Linking Design: Utilizing a Single Set of Anchor Items with Fixed Common Item Parameters during the Calibration Process*

ABSTRACT

This study has explored the relative merits of a potentially useful IRT linking design: Utilizing a single set of anchor items with fixed common item parameters (FCIP) during the calibration process. An empirical study was conducted to investigate the appropriateness of this linking design. A parameter recovery study was performed to examine the robustness of FCIP under the situation of large standard errors in the item difficulty and guessing parameters. Comparison of results to those produced by the characteristic curve method (CCM) was pursued. Based on the empirical portion of this study, ability estimates calibrated from this linking design are very consistent, except for students with extreme (especially low) ability under the CCM equating method. Item parameter estimates calibrated from this linking design are also very consistent, except for guessing parameters under the CCM equating method. Based on the results from the simulation portion of the study, this linking result can produce very precise and stable parameter estimates.

Key Words: Linking Design, Linking Item Parameter, Test Equating, Item Response Theory (IRT); Characteristic Curve Method (CCM).

---

* The authors would like to convey special thanks to the former Director of the Department of Test Development and Administration, Dr. Leroy Tompkins, for his suggestions on this linking design, Director, Dr. Valeria Ford, for her continuous support on this project and Mrs. Florence Dichter for her careful proofreading on this paper.

I. Introduction

A. Background of IRT Equating Methods

Various item response theory (IRT) models have now been routinely used to construct a large item bank with all items on the same metric, owning to the nice property that the metric thus calibrated is assumed to be invariant under a linear transformation (Lord, 1980). After an item bank has been carefully constructed, the abilities of a different group of examinees can then be calibrated on the same numerical scale, even though their abilities may have been estimated from a different subset of items.

Practically speaking, developing an item bank is complicated and time consuming, involving such processes as pilot test construction, item calibration and item linking. In addition to striving for improvement in item calibration techniques (see Baker, 1992), test developers should also pay attention to creating a simple and feasible linking design. Ideally, it should hold the features of reducing the efforts in data collection and item linking, and yet produce equating results with minimal equating errors.

The concern of linking is to develop an item pool from several pilot tests administered to different groups of examinees by calibrating the various item parameters onto a common metric (Vale, 1986). One such approach is to calibrate tests separately in different groups and then to convert all item parameter estimates to the same scale through a linear transformation. Several other methods, such as the characteristic curve method (CCM) (Stocking & Lord, 1983) and the minimum chi-square method (Divgi, 1985) are also available. As anticipated, the linking parameter estimates produced by different transformation methods can be different. Besides, these types of linking transformation methods ignore the effect of the guessing parameter estimates on the metric conversion among tests. The reasons given are usually two-fold. First, since the guessing parameter is measured on the probability axis, no transformation needs to be applied. Second, while the minimum chi-square method (Divgi, 1985) can take the information of the guessing parameter and its covariances with other item parameters (i.e. discrimination and

difficulty) into account while estimating the transformation coefficients, it is not safe because the guessing estimate is usually unstable and its covariances may be too large.

Nevertheless, there are two equating methods that can incorporate the information of the guessing estimate for metric transformation. The first one is the concurrent calibration method, which combines test data sets by treating those items not taken by any particular group as "not reached items" during compilation (Hambleton, Swaminathan & Rogers, 1991). Another method is to fix common item parameters (FCIP) among tests during the calibration process (Mislevy & Bock, 1990).

The concurrent calibration equating method satisfied Mislevy and Bocks' (1982) requirement that "the information needed for a proper link is found in not just the item parameter estimates and their standard errors, but in the matrix of correlations among the estimates as well" (p. 15). Since this method makes complete use of the available information and may potentially remove some equating errors produced by the inaccurate transformation functions that are used to equate the two tests, it appears that this linking method may produce a more stable equating result than many other linking methods did (see Kim & Cohen, in press; Kim & Cohen, 1997). This method and the FCIP, however, may encounter problems when equating two or more tests which were taken by heterogeneous groups (discussed in the following section).

FCIP, while holding some of the same properties as the concurrent calibration method, may also produce more stable linking results than the linking transformation methods (e.g. remove some equating errors produced by the inaccurate transformation functions, take the guessing parameter into account for metric conversion). Moreover, it may alleviate several practical problems (to be discussed below) that occurred in the concurrent calibration method.

Of the numerous data collection designs, the anchor-item data collection method is often used in many large standardized testing programs. However, in the case of constructing more than two pilot tests, there exist several alternative data collection designs (refer to Vale, 1986). Among these, it is much easiest to create pilot test forms with a single set of common items, thereby allowing test developers more time to focus on designing a sound set of anchor test items.

When the FCIP linking method is to employ a single set of common items across all tests, the conspicuous feature of this linking design is that common items are administered to the accumulated group of test takers so that more stable common parameter estimates, especially the guessing parameters (discussed in section two below), can be expected. Consequently, when these common parameter estimates are used as the anchor of the measurement scale, more accurate equating results can be expected.

B. Statements of Research Question

The utilization of a single set of anchor items under FCIP has been implemented in a school district on the Eastern shore. Several item banks (see section three below) have been developed using this linking design. The present study investigates the appropriateness of this linking design. Specifically, this study concerns itself with whether this linking design produces more stable equating results, especially for the guessing and the ability parameter? When items are relatively easy, the estimates of the location and the guessing parameters may be relatively unstable (discussed in section two below). Can this linking design minimize this problem? To answer these questions, tests with relatively easier items will be generated by centering the ability parameter one standard deviation above the mean item difficulty. The performance of the FCIP linking method under the above condition will be evaluated by means of a simulation study.

In short, the main purpose of this study is to illustrate the efficiency of this linking design as well as to formally examine the robustness of the FCIP linking method under the

situation of large standard errors in the item difficulty and guessing parameters. The robustness issue will be examined by analyzing a real data set as well as by conducting a simulation study. Comparison of results to those produced by the CCM linking method will be pursued.

Presented are a brief review of the issues of item linking and data collection in section two, a description of the methodology in section three, the results and discussions in section four, and the conclusions in section five.

II. Review of Issues Related to Item Linking and Data Collection

A. Item Linking

In IRT, numerical estimates of the item parameters depend upon the ability (θ) scale (Baker, 1992). Although, the ability scale is usually standardized to have a mean of zero and a standard deviation of one in "any" data set being analyzed, the original (not standardized) ability scale is different from one data set to another. Thus, when the "same" set of test items is administered to two different groups and the resultant response data are calibrated separately, the two sets of item parameter estimates are usually different because they refer to different underlying ability scales. This problem can be resolved by using one of the linking methods reviewed below. However, it should be pointed out at the forefront that the common scale thus constructed is still arbitrary and is limited. Its interpretation must be carried out with caution.

1. Determination of Equating Constants: Characteristic Curve Method

Under the three-PL logistic model (see Baker, 1992; Hambleton & Swaminathan, 1985; Mislevy & Bock, 1990), the probability, P_{ij}, of a correct response to the i^{th} item for the j^{th} examinee with ability θ_j is given by:

$$P_{ij}(\theta_j) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}} \qquad (1)$$

where a_i is the item discrimination, b_i is the item difficulty, c_i is the lower asymptote parameter (also known as the guessing parameter), and D (usually equal to 1.702) is a scaling factor. A two-PL model is attained if the guessing parameter c_i is constrained to zero for all items in (1) above. A one-PL model (or known as a Rasch model) is a restricted form of the two-PL model by further constraining the item discrimination index a_i to be identical or equal to one for all items.

It is well known that item parameter estimates from two separate calibration procedures will differ in terms of their origins and units of measurement (Cook & Eignor, 1983; Kolen &

Brennan, 1995). Yet it is possible to equate the two test forms with simple linear equations by using a set of common (or anchor) items. The fact that IRT scales can be transformed via a linear function only holds when the model selected fits the data. If there is significant lack of fit, this property will no longer be valid.

Let (b_B, a_B) be the difficulty and discrimination parameters for the based group, and (b_E, a_E) be the corresponding parameters for the equated group. The relationship between the parameters from the two test editions can be obtained by fixing the metric of θ in each group. Equating can be achieved as follows: (see Hambleton & Swaminathan, 1985):

$$b^*_B = \alpha b_E + \beta \quad (2)$$

$$a^*_B = a_E/\alpha, \quad (3)$$

where the superscript * represents the transformed values from the equated group to the base group, α is the slope, and β is the intercept.

When the pair of θ's, b's and a's are known, the equating constants (α and β) can be determined by means of a simple linear transformation. Yet in practices, only estimates of the parameters are available. As a result, several alternatives (see e.g. Divgi, 1985; Linn, Levine, Hastings & Wardrop, 1981; Loyd & Hoover, 1980; Stocking & Lord, 1983) have been developed for estimating the α and β parameters. Among these approaches, the minimum chi-square method may (Divgi, 1985) produce stable linking parameter estimates as it takes into account the most information available from the item difficulty, the item discrimination and the item parameter variance-covariance matrix. Another possibility is the characteristic curve method, or CCM, which matches the characteristic curves between the base and the equated group (Stocking & Lord, 1983). It is "logically" better than other approaches if the primary concern of a test practitioner is the equivalency of the true scores between two groups of examinees. Part of the reason for the popularity of this method can be attributed to the availability of a freely distributed computer program called EQUATE 2.0 (Baker, 1993). Since the CCM equating

method is frequently used in both research and practical test settings, a more detailed review is presented below.

The basic principle behind CCM is to minimize the squared differences between the two true scores derived directly from the two sets of common-item parameter estimates for N arbitrary ability points, θ's. In symbols, it is given by:

$$F = \frac{1}{N}\sum_{i=1}^{N}[T(\theta_i) - T^*(\theta_i)]^2 \qquad (4)$$

where

$$T(\theta_j) = \sum_{i=1}^{L} P_i(\theta_j, a_i, b_i, c_i) \qquad (5)$$

and

$$T^*(\theta_j) = \sum_{i=1}^{L} P_i(\theta_j, a_i^*, b_i^*, c_i). \qquad (6)$$

In equations 5 and 6, $T(\theta_j)$ refers to the true score of an examinee with an ability θ_j on a set of common items in the base test (B), and $T^*(\theta_j)$ refers to the true score of an examinee with the same ability θ_j on the same set of common items, in which the parameters of the common items have been transformed from the equated group to the base group (see equations 2 and 3). Finally, L denotes the number of common items.

After obtaining the equating constants, the item parameters derived from the different subtests can be transformed to the same scale. Of course, these equating constants can also be used to convert between two metrics of ability estimates. The equation for ability metric transformation is the same as in equation 2, except that the b parameter should be replaced by the ability parameter. The performance of CCM in producing transformation coefficients is well documented in the literature (Baker & Al-karni, 1991; Baker, 1996).

2. Concurrent Calibration

Item linking for different test forms can also be accomplished within a single calibration run by the concurrent calibration method. The item and ability parameter estimates calibrated by such an approach are automatically on a common scale (Mislevy & Bock, 1990). Popular computer software packages such as BILOG (Mislevy & Bock, 1990) provide this nice feature. The most conspicuous feature of this method and the FCIP may minimize the impact of sampling fluctuations on estimating the guessing parameters. The reasons for this is given below.

Unless a large number of examinees or tight priors around the true parameter values are used, the guessing estimates will be unstable (van der Linden & Hambleton, 1997). Generally speaking, only low-ability groups are informative in the estimation of the guessing parameter (Stocking, 1993). When there are few low-ability examinees available to estimate the guessing parameter of an easy item, the corresponding standard error can be very large. In addition, the large covariance between the guessing parameter and the difficulty parameter can "cause this uncertainty to move partially to the estimate of location" of the item difficulty. This effect is lessened somewhat for more difficult items (Thissen & Wainer, 1982). Fortunately, when the sample size is relatively large, this problem can be minimized. Since the concurrent calibration and the FCIP linking methods calibrate the common items parameters based on the accumulated group of examinees from several tests, they may thus take advantage of a larger sample size. Estimation of other parameter estimates may profit in turn once the anchor metric defined by the common item parameters is more stable.

There are, however, potential problems with the concurrent calibration method when the test data are originated from groups with extreme differences with respect to the locations and variabilities of the ability distributions (see Kim & Cohen, 1997). This issue deserves further investigation. Besides, test practitioners are faced with two practical problems. Firstly, coding for the "not reached items" can be very tedious when more than two test forms are equated simultaneously. Secondly, this linking method is not very handy to calibrate new items into an

existing scale because it requires a combination of the new dataset(s) with the original test dataset(s) that might already be unavailable or stored in the database warehouse.

3. Fixed Common Item Parameter (FCIP) during the Calibration Process

Earlier, Mislevy and Bock (1990) indicated that: "By specifying tight priors on selected item parameters, the user may hold these values essentially fixed while estimating other item parameters. This feature is useful in linking studies, where new test items are to be calibrated into an existing scale without changing parameter values for old items" (pp.2-6, 2-7).

Although the above principle is mainly used for adding new items into an existing item bank, it can also be used to develop new item banks if the common-test item estimates are precise and stable. The FCIP method abides by this principle according to the following procedures. First, it estimates the common item parameters with examinees accrued from all groups of test takers assigned the same set of anchor items. Second, it holds the common parameters fixed at those values obtained from the first step, and also specifies the standard errors of the common item parameter estimates close to zero during the calibration process.

FCIP, in essence, is equivalent to the concurrent calibration method except that the FCIP does not take the information of item parameter variance-covariance into account on the metric conversion among tests. The FCIP might also encounter the problem as the concurrent calibration may occur under the circumstance that tests are administered to groups with extreme differences in location and variability of ability. Based on the empirical experience of the first author, FCIP may encounter several problems especially in a vertical linking situation. First, FCIP may produce extreme values of item parameter estimate (e.g., slope =52.984; SE=15.364). Second, some common item parameters, especially the discrimination parameters, can not be fixed at the numeral values as desired. Under this situation, the practitioner may need to free these item parameters in order to reach the convergent criterion. Third, the iterative procedures for finding some of the item parameter estimates may fail. Lastly, the program may fail to reach

the convergent criterion as set up by the practitioner. However, these situations are uncommon in empirical settings.

Some advantages of FCIP include its flexibility in being adapted to a variety of data collection methods (see Vale, 1986), and that it may produce less equating errors of item parameters, especially the guessing parameter, than other linking transformation methods. Nevertheless, the knowledge about FCIP is rare and needs to be further investigated.

B. Data Collection Design

There exists several data collection methods for developing an item bank (see Petersen, Kolen & Hoover, 1989; Vale, 1986). Two widely used approaches in developing a large item bank with anchor-test-item are as follows. The first one makes use of several sets of common items. For instance, Form 1 and Form 2 may use the same set of common items for item linking, whereas Form 2 and Form 3 use another, etc.. The second method is to employ a single set of common items across all pilot tests.

Given the same conditions, the first design above will of course produce more test items than the second one. However, test developers will need to invest more time and energy in creating pilot test forms and linking items according to the first design. For example, if six test pilot forms are created, five sets of anchor test items are needed to begin with. As delineated elsewhere, the characteristics of anchor test items can affect the results in test equating studies (see Brennan, 1987; Cook & Petersen, 1987). Locating simultaneously five sets of sound anchor test items is definitely not easy. This problem is further compounded with the practicality issue when item linking is actually carried out. It will require a lot of steps to link five sets of anchor items when, say, the CCM approach is used.

In contrast, the utilization of a single set of anchor items across all pilot tests is much easier. Besides, this method, due to its flexibility, also facilitates the adaptation to almost all existing linking methods.

C. An IRT Linking Design: A Single Set of Anchor Items with FCIP Linking

A linking design consists of two components, namely, data collection methods and linking methods. As discussed above, there are several data collection methods and linking methods available. As a result, there exist a variety of linking designs.

The design using a single set of anchor items with the FCIP linking method can be a potentially useful design with the following advantages:

(a) The quality of a single set of common items is easier to control.

(b) Several linking methods can be applied to the data thus collected.

(c) If the number of test takers accrued from all groups is large enough or approaching that of the population of research interest, the common item parameters calibrated from the accumulated sample size may be close to the true item parameters (Hambleton & Cook, 1983; Swaminathan & Gifford, 1983).

(d) As more examinees are available, more information can be provided during the process of estimating common item parameters. With more stable common parameters, FCIP can then result in less equating errors.

(e) The FCIP takes the guessing parameter into account for the metric conversion among tests.

(f) The FCIP may remove some equating errors produced by the inaccurate transformation functions

(g) The ability and various item parameters can be estimated simultaneously.

Consider the example that six pilot test forms are created and administered to 1000 subjects per test form. The common anchor test items are then taken by a total of 6000 subjects. With this larger sample size, precise and stable item parameter estimates are obtained even though the test length may be quite small, for example 15, (Harwell & Janosky, 1991).

Since this linking method relies heavily on the level of precision of the estimates of the common item parameters, apparently this linking design is more appropriate for equating

multiple tests rather than just two tests. The practical issue of whether this linking design can be effectively implemented to equate two tests will also be investigated in this study.

III. Methodology

Since quite a number of research issues are being dealt with in this study, an overview will first be given followed by a detailed description of the research design for each portion of the study.

A. Overview

In this section, an empirical study that investigates the appropriateness of FCIP will first be described. Its results will then be compared to those produced by the CCM linking method. The concern about the effectiveness of implementing FCIP to only two test forms will also be carried out.

A parameter recovery study will then be performed to investigate the robustness of the FCIP linking results when the standard errors of the item difficulty and guessing parameters of the easier items are relatively large. Technically, this condition was attained by generating item response matrices from a test administered to a high-ability group of examinees (ability centered about 0.84 standard deviation above the mean item difficulty). This portion of the study is conducted by simulation while imitating the testing situation in the empirical study. Towards this end, item parameters of the first three test forms are selected from the empirical study. In addition, three types of underlying high-ability distributions, namely the positively-skewed, the normal, and the negatively-skewed, will be generated. Each test form will be randomly administered to only one of the three ability groups. Twenty replications of test data will be generated for each test form and shape of ability distribution combination. The results of linking item parameters and equating ability scores will then be analyzed. Comparison of results to those produced by the CCM linking method will be pursued.

B. An Empirical Study

1. Sample Size

The sample sizes of examinees taking each pilot test are listed in Table III-1. The accumulated sample size from the six groups thus amounts to a total of 8,357 students, almost

equivalent to the target Grade Two Students population size. The cluster sampling unit being used is school. Moreover, the test developers also attempt to make the ability level of each sample similar by taking into consideration the past records of student performance in each school on a Math criterion-referenced test.

Table III-1.
The Sample Sizes for Each Pilot Test

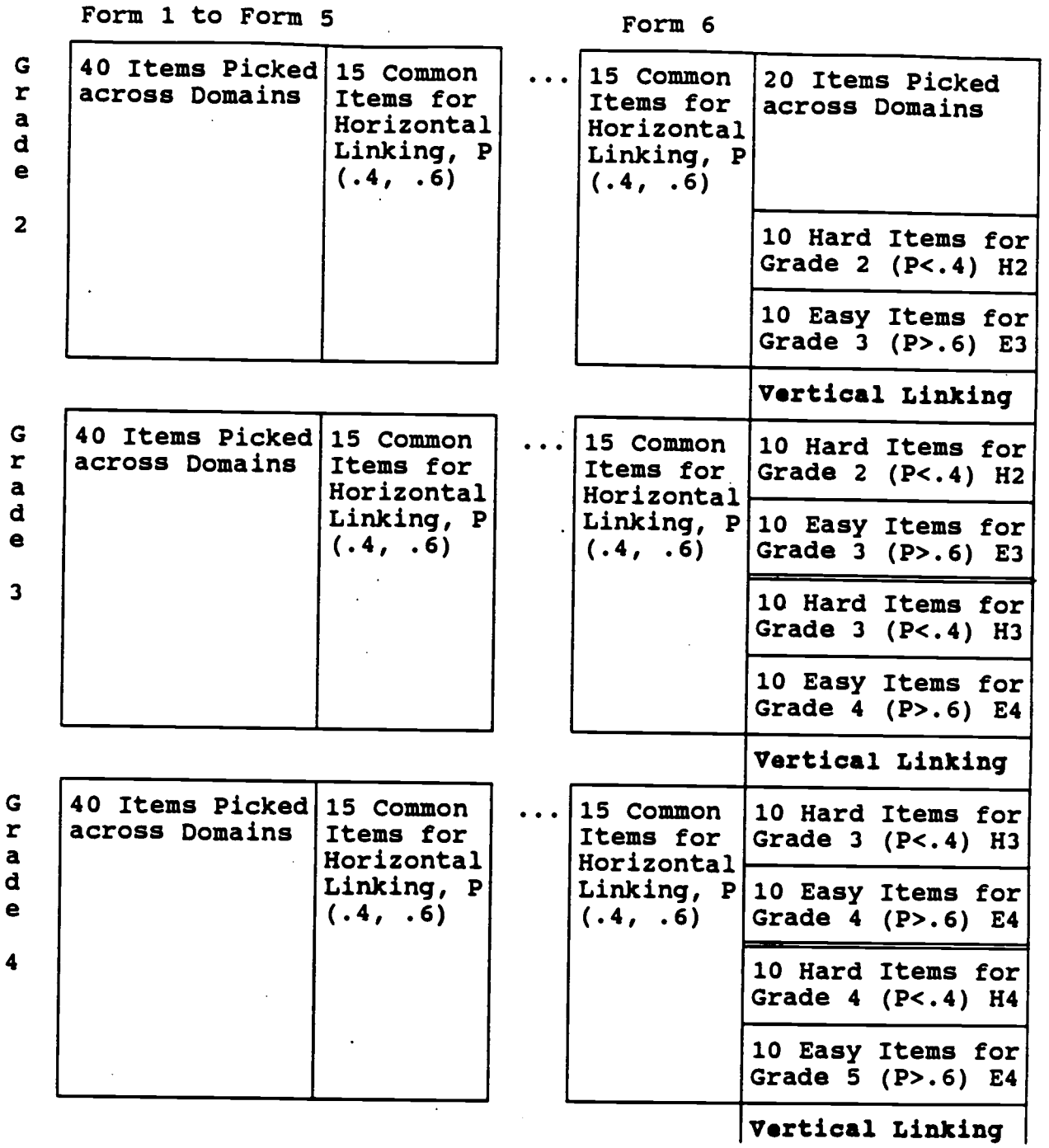
	Form_1	Form 2	Form_3	Form_4	Form_5	Form_6	Population
Sample Size	1249	1451	1443	1232	1353	1629	8357

2. The Construction of Pilot Test Forms

The test specification tables used are based on the "Elementary Mathematics Outcomes" (1992) standards. Each test was constructed in a paper-and-pencil multiple-choice format with four options. Six pilot test forms, all using a single set of common items, are created for the purpose of developing a mathematics item bank. As seen from the top part of Figure 1, each test form (from 1 to 5) consists of 40 unique items and 15 common items which were used for horizontal linking. The 15 common items are constructed as a miniature test, which resembles the overall test by evenly picking moderately difficult items from eight domains. The percent of correct values, P's (i.e. the item difficulty index from classical test theory or CTT), of the common items are set between 0.4 and 0.6 (see Figure III-1). The reason for choosing the P values based on CTT, rather than the b values based on IRT, is simply because the b values have not been produced at that point yet.

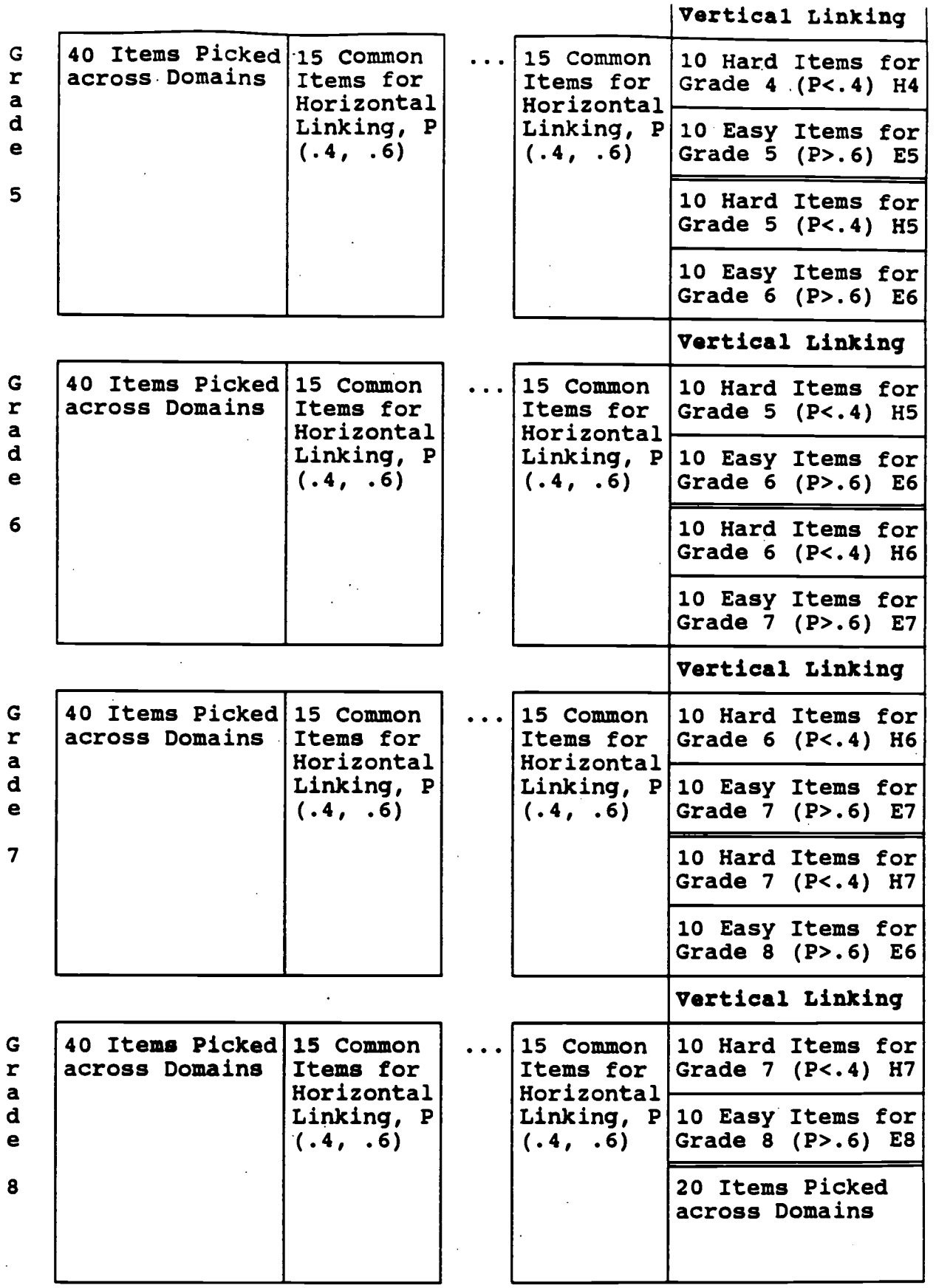
The primary concern of Form Six is for vertical linking which is beyond the purpose of the present study and will not be discussed here. Basically, Figure III-1 is a test blueprint including horizontal equating within each grade and vertical equating across grades from Grade 2 to 8. So far, the mathematics item bank for each grade has been successfully developed.

Figure III-1 A Plot of Developing Item Pools by Horizontal and Vertical Linking



(Continued on next page)

Figure III-1 continued



3. Determining the Item Response Model and Estimating Item Parameters

The three-parameter logistic model is selected to fit the data. The primary reason for choosing the three-PL is because the students have been encouraged to pick an answer they deem most appealing if they are unable to locate their answer. Moreover, the large sample size used in this study conforms to that required by the three-PL.

Item parameters in each subtest are calibrated by BILOG (Mislevy & Bock, 1990). Performance assessment of this software can be found in Ackerman (1987), Algina (1986), Baker (1990), Mislevy and Bock (1982, 1984), Harwell & Baker (1991), Mislevy & Stocking (1989) and Yen (1987). Two options in BILOG, however, deserve further explanation.

The first one is the FREE option, which when adopted, will instruct the program to empirically estimate the θ distribution of the respondents. Otherwise, the default is to assume the ability parameter to be distributed as a unit normal. In the present study, this option is invoked because, for a test length equals 55, the empirical posterior can be quite accurately estimated. The second one is the FLOAT option. If this option is adopted, the means of the item parameter prior distributions will be estimated along with the item parameters (see Mislevy & Bock, 1990). Otherwise, the means of the item parameters distribution will be fixed at their default values during the estimation process. In this study, FLOAT is invoked.

In addition, the maximum number of EM cycles is set at 100, the maximum number of Newton-Gauss iterations following the EM cycles is set at 20, and so is the number of quadrature points for numerical integration. The expected a posteriori (EAP) method is chosen to estimate the ability scores (see Bock & Aitkin, 1981; De Ayala, Schafer & Sava-Bolesta, 1995, for detailed discussion).

4. Item Fit Measure and Item Dregging

Eliminating test items by item fit statistics should not be automatically determined by any computer program. This is because most item-fit statistics are not reliable in the first place (see

Hambleton & Murray, 1983; McKinley & Mills, 1985; Reise, 1990; Rogers, 1987; Tam & Li, 1997; Yen, 1981). The indices derived from BILOG are only used as a reference in evaluating items in the present study. Eventually, some items are removed based on a combination of criteria, such as the item fit statistics, teachers' comments etc..

5. Linking Procedure

(1). CCM Procedures

The computer program EQUBANK (Li, 1997), which was written in MATLAB (The MathWorks Inc, 1995), is used to perform the CCM linking. This program can produce exact numerical values of the transformation coefficients as the EQUATE program under the same condition. There is no limit to the number of item parameters to be transformed in EQUBANK. With this program, the item parameter estimates of Form 1, to Form 5 are transformed to the same origin and measurement unit of Form 6. The equating coefficients of α and β from various Forms to Form 6 are listed in Table III-2. After the metric transformation, each item parameters are put on the same scale except for the numerical values of the 15 common item parameter estimates, which will be a little bit different from one Form to another. The final estimates of the common item parameters are obtained by taking the average of these six sets of 15 common item parameter estimates.

Table III-2
Equating Coefficients Computed by the EQUBANK Program

From	To	α	β
From Form 1	to Form 6	1.0075	-0.1820
From Form 2	to Form 6	0.9876	-0.1664
From Form 3	to Form 6	0.9883	-0.0643
From Form 4	to Form 6	1.0424	-0.0016
From Form 5	to Form 6	0.9926	-0.0392

(2). FCIP Method

Under FCIP, the 15 common items across the six pilot tests are first calibrated from the total population. The numerical values of the 15 common item parameter estimates are then fixed at the values thus derived, while estimating the rest of the items in a BILOG run. After that, item linking is accomplished automatically.

As regards the effectiveness of implementing FCIP to only two test forms, another FCIP linking method is conducted by fixing the numerical values of the common item parameter estimates to those based on 2000 cases randomly sampled from the population. The reason for sampling 2000 examinees is because 2000 students will take the common test items when only two test forms are administered to 1000 students per test form.

(3). Metric Conversion between the Item Bank Linked by FCIP and the Item Bank Linked by CCM

The metrics of the final linking estimates from FCIP and from CCM are different but interchangeable. In order to make comparisons possible, an additional EQUBANK run is conducted to transform the metric derived from the CCM method into that obtained from the FCIP method.

6. Data Analysis

Two analyses will be pursued for this part of the study. Firstly, the degree of association between the parameter estimates from CCM with those from FCIP will be assessed. Secondly, the residuals between each pair of parameter estimates (those from CCM and FCIP) will be analyzed by descriptive statistics.

C. Recovery Study by Simulation

Essentially, the test data generated for the recovery study is obtained from simulating the testing situation as in the empirical study. The main purpose of this simulation study is to

investigate the robustness issue of FCIP. The final results of linking item parameters and equating ability scores will be assessed by comparing them to their corresponding true values.

1. The Simulation of Test Data

Three simulated tests forms are generated for this part of the study. The item parameters are selected from the first three tests in the empirical study. The descriptive statistics of the item parameters used to simulate these three tests are presented in Table III-3. On the whole, the three tests have similar item characteristics. The numerical values of the item parameters are all based on an ability distribution that is distributed as a standard normal. Thus the large negative b values can be interpreted as relatively easy items. The average b values for the common items is, relatively speaking, less than those of the rest of the test items across the three test forms. This condition is similar to the vertical linking situation when FCIP is used for parameter calibration for each single test.

Table III-3
Descriptive Statistics of the Item Parameters for the Three Simulated Tests

		a		b		c	
	TL	Mean	SD	Mean	SD	Mean	SD
Form_1	53	0.909	0.243	-0.442	0.987	0.161	0.062
Form_2	52	0.978	0.257	-0.438	1.095	0.174	0.080
Form_3	53	0.871	0.236	-0.378	1.038	0.186	0.076
Anchor	15	0.851	0.220	-1.080	0.712	0.099	0.055

2. Ability and Sample Sizes

High-ability examinees will be simulated. The reason behind this decision is to purposely generate larger standard errors of the b and c parameter estimates for easy items when the number of low ability examinees are few. The intention here is to find out if FCIP can handle this problem better when compared to such convention transformation methods as CCM. To put it another way, the concern here is to find out how robust FCIP is when the standard errors of estimating b and c can be quite large.

Referring to the number axis, results from the first empirical study indicate that the ability distribution is located about 0.42 standard deviation to the right of the item difficulty location. In order to generate a higher ability group, this distance is doubled so that the ability distribution will center at about 0.84 standard deviation above the mean item difficulty. So far as horizontal linking is concerned, even though the ability locations of various groups taking different test forms are quite close in this portion of the study, yet the shapes of the ability distributions may make a difference. Thus three types of examinee ability groups that can occur in real life will be generated. The first one is normal distribution. The second one is a positively-skewed distribution which is characterized by a chi-square distribution with eight degrees of freedom (see Seong, 1990). The skewness of this distribution is 1. The last one is a negatively-skewed distribution which is the mirror image of the positively-skewed distribution mentioned earlier. All three shapes of high-ability distributions are standardized to a mean of 0.42 and a SD of 1.

3. Test Dataset Generation

Each simulated test will be administered to only one of the three groups. The computer program GEN3PL01 (Li, 1996) is used to generate the test data. This program was also used in the study by Tam & Li (1997). This kind of data generation will reflect the "real" and "practical" testing settings. Twenty replications of test data will be generated for each test form and shape of ability distribution combination.

4. Calibration Analysis and Linking Procedure

Both calibration and linking are conducted in the same way as in the empirical study. Within each replication using CCM, the metrics of item parameter estimates from Form One and Three are converted into the anchor metric of the item parameter estimates from Form Two, which is arbitrarily assigned to the normally distributed ability group. Finally, this metric is transformed into the metric defined by the true parameters.

Within each replication using FCIP, the common metric of the item parameters of the three tests is defined by that of the common parameter estimates. This common metric is then transformed into the metric defined by the true parameters via EQUBANK.

However, when the data are generated by examinees with negatively-skewed high-ability distribution, one or more than one (or two to five) common item parameter(s) is/are unable to be fixed at the specified values during the calibration process. For instance, suppose that the item discrimination and difficulty of Item 2 should be fixed at 0.885 and -2.085, respectively. Unfortunately, they come out 3863.273 and -1.00 respectively at the end of the calibration and the program fails to reach the convergent criterion chosen by the users. This may happen in fifteen out of the 20 replications. When this situation occurs, those common items causing the problem are set to be "FREE" with the rest of the "non-common" test items. The above problem may affect the metric conversion especially if too many common item parameters can not be fixed.

5. Evaluation

Twenty replications are simulated. Once the estimated parameters are obtained, the accuracy of the parameter estimates is assessed by using the BIAS and the root mean square error (RMSE) criteria. They are computed as follows (see Skaggs & Lissitz, 1988):

$$\text{BIAS}(H) = \frac{\sum_{i=1}^{p}(\hat{H}_i - H_i)}{p} \qquad \text{(III-1)}$$

and

$$\text{RMSE}(H) = \sqrt{\frac{\sum_{i=1}^{p}(\hat{H}_i - H_i)^2}{p}}, \qquad \text{(III-2)}$$

where H_i is the true parameter, $\hat{H}_i$ is the corresponding estimated parameter, and p is the total number of replications.

IV. Results and Discussion

A. An Empirical Study

1. Comparison of the Ability Estimates Under FCIP and CCM

The main concern in the empirical study is the residuals between the ability estimates from FCIP and CCM.

Using the equating coefficients listed in Table III-2, CCM estimated student abilities from the six test forms separately, which were then transformed into the anchor metric defined by the item parameters in Form Six. Meanwhile, student abilities were also estimated by FCIP with the common items across all six test forms being fixed during the calibration process. The correlation coefficient between the two sets of ability estimates is 0.998. The plot of the equated abilities by CCM against those by FCIP is presented in Figure IV-1 below.

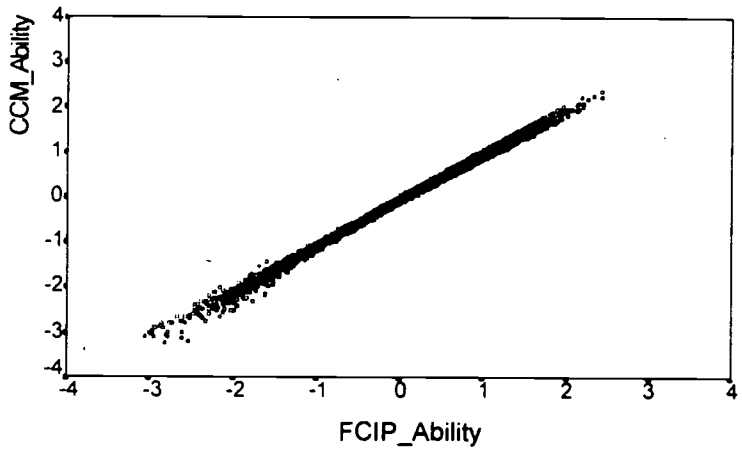


Figure IV-1: The Relationship between the Ability Estimates by FCIP and CCM

The CCM equated ability estimates were then converted into the metric defined by FCIP using the item linking transformation coefficients of α =0.9618 and β= 0.0957. Afterwards, residual for each student was computed by taking the difference between the corresponding estimates from FCIP and CCM. The average residual of all students is -0.037, with a standard deviation of 0.055. The minimum and maximum residual values are -0.372 and 0.468,

respectively. The plot of residuals against the FCIP equated ability estimates is presented in Figure IV-2. This plot provides a clear picture as to the nature of consistency between the ability parameter estimated by the two methods. Apparently, inconsistency is localized in the region of extremely low ability estimates ($\theta < -2.0$). The rest of the student ability estimates, about 99% of them, from the two equating methods are quite consistent.

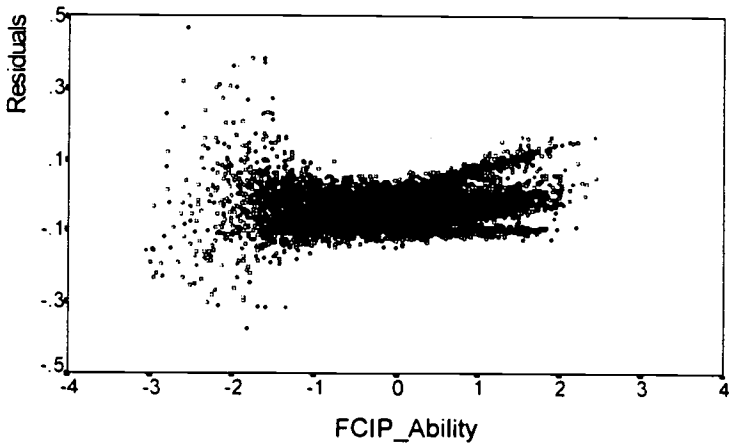


Figure IV-2: The Plot of Residuals against the Ability Estimates by FCIP

2. Comparisons of Item Parameters Linked by FCIP and CCM

Similar comparisons of item parameters linked by FCIP and CCM were also performed. Using the equating coefficients listed in Table III-2, CCM calibrated the item parameters separately from the six test forms and then transformed them into the anchor metric defined by the item parameters in Form Six. On the other hand, FCIP estimated the item parameters by fixing the common items across all six test forms during the calibration process. The Pearson correlation coefficients of the item discrimination, difficulty and guessing parameters between the two sets of estimates were found to be 0.986, 0.996 and 0.799, respectively. Figures IV-3, IV-4 and IV-5 are plots of relationship between the two sets of item discriminations, item difficulties and guessing parameters, respectively. These figures show that the relative positions of the pairs of item parameters linked by FCIP and CCM are very consistent, except for the guessing parameter estimates.

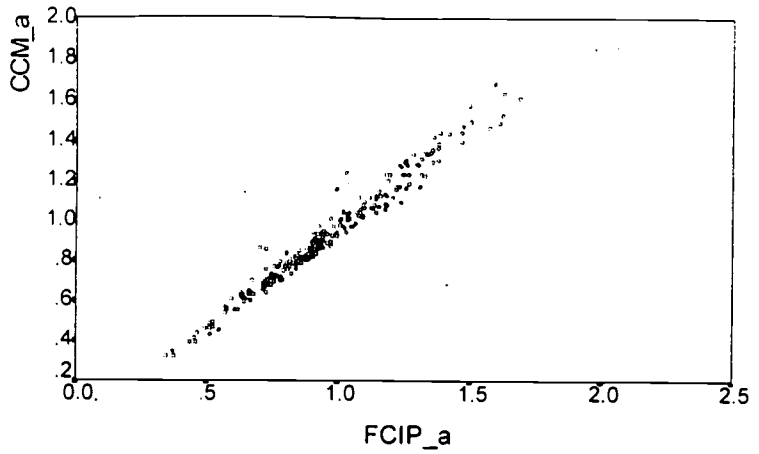


Figure IV-3: The Relationship between the Item Discrimination Estimates by FCIP and CCM

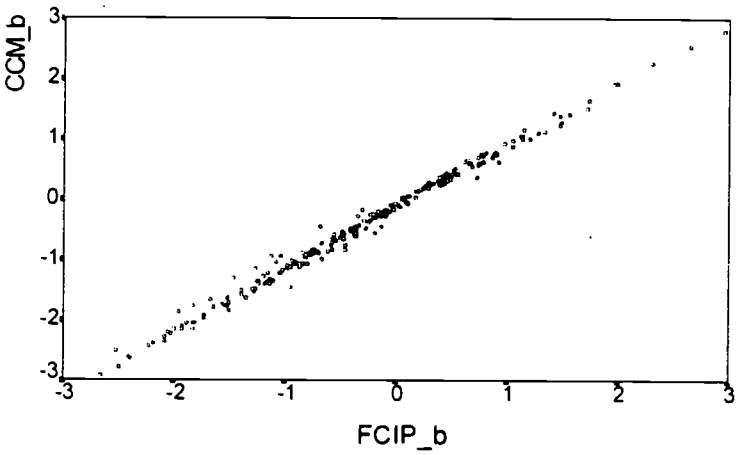


Figure IV-4: The Relationship between the Item Difficulty Estimates by FCIP and CCM

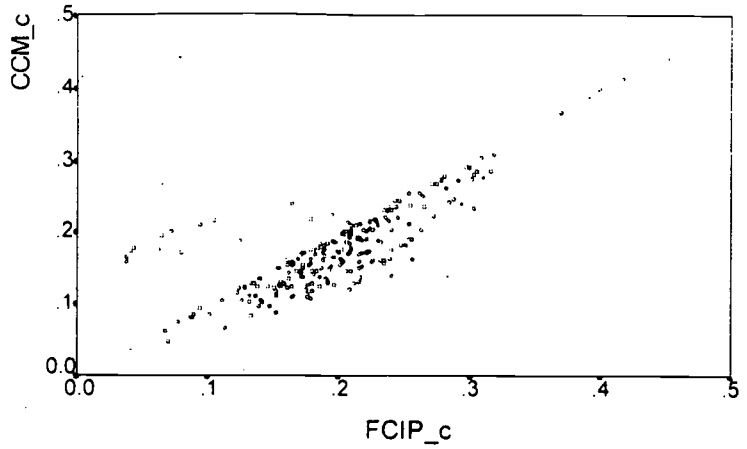


Figure IV-5: The Relationship between the Guessing Parameter Estimates by FCIP and CCM

An additional EQUBANK run was conducted to transform the metric of a set of item parameter estimates defined by CCM into that defined by FCIP. The transformation coefficients of α and β were the same as those used for equating the two ability metrics. Residuals were then computed by taking the differences between the linking parameters from FCIP and those from CCM. Descriptive statistics of the residuals are presented in Table IV-1. The average residuals are 0.06, -0.008 and -0.020, and the standard deviations are 0.05, 0.039 and 0.039 for the item discrimination, difficulty and guessing parameter, respectively. The plots of residuals versus the corresponding FCIP's discrimination, difficulty and guessing parameter estimates are presented, respectively, in Figures 6 to 8. Based on the residual analysis, item parameter estimates equated by these two approaches are very consistent. However, items with high discrimination, low difficulty or low guessing parameter under FCIP are potentially inconsistent from those from CCM.

Table IV-1

Descriptive Statistics of the Residuals: The Differences between the Linking Parameters from the CCM and Those from the FCIP (Number of Items in the Item Bank =239)

	Mean	SD	Skewness	Kurtosis	Minimum	Maximum
a	0.006	0.050	1.440	3.922	-0.098	0.260
b	-0.008	0.093	0.373	2.770	-0.336	0.334
c	-0.020	0.039	1.832	5.639	-0.099	0.135

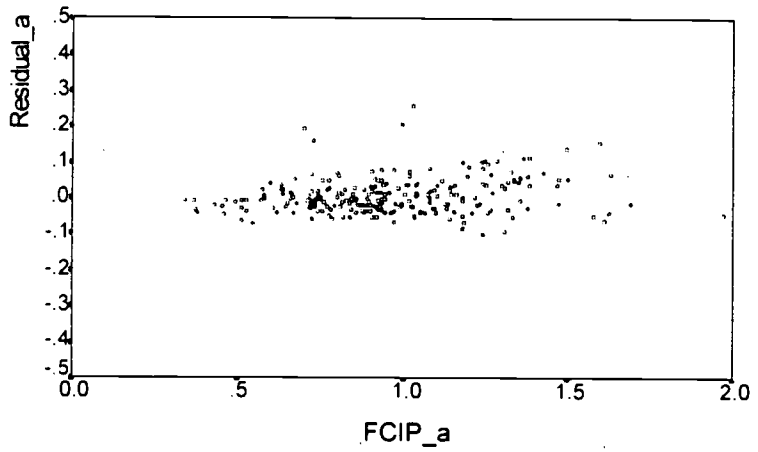


Figure IV-6: The Plot of Residuals_a against FCIP's a Parameter Estimates

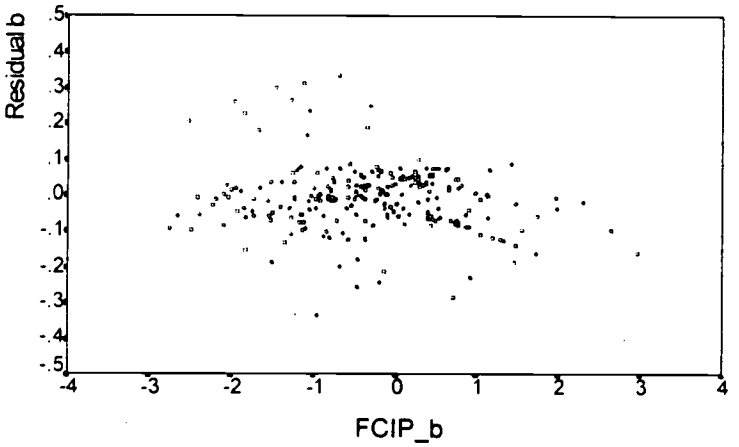


Figure IV-7: The Plot of Residuals_b against FCIP's b Parameter Estimates

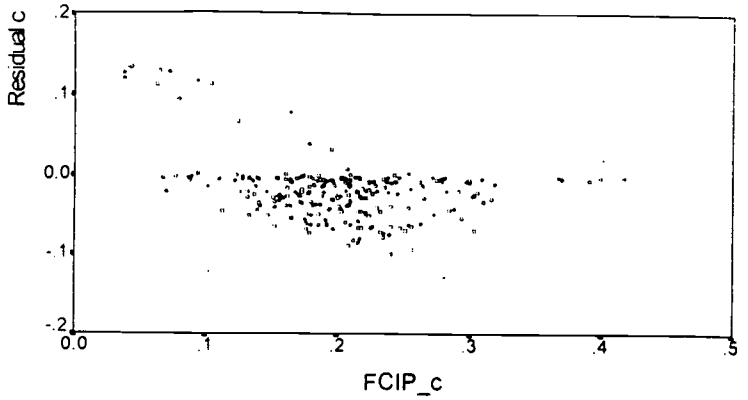


Figure IV-8: The Plot of Residuals_c against FCIP's c Parameter Estimates

3. The Empirical BIAS and RMSE of the Guessing Parameter Estimates

Table IV-2 shows the impact of sampling fluctuations on estimating the guessing parameters. Since the guessing parameter estimates are on the probability metric, they should not be equated. Logically, the guessing parameter estimates of the 15 common items should be fairly close across the six samples. Yet in reality, they were affected by sampling. For instance, the guessing parameter of item 3 in Form 1 was 0.154; but 0.300 in Form 4. In contrast, the guessing parameter derived from the population was 0.163. This particular problem can not be dealt with by the CCM approach. The empirical BIAS and RMSE values for each common item were calculated using Equations III-1 and III-2. They are presented on the right hand side of Table IV-2. It appears that the guessing parameters are often overestimated.

Table IV-2

The Common Item Guessing Parameter Estimates Calibrated from the Six Different Tests Taking by Different Samples and the Corresponding Empirical BIAS and RMSE

ID	Form1	Form2	Form3	Form4	Form5	Form6	Population	BIAS	RMSE
01	0.271	0.223	0.203	0.200	0.245	0.215	0.195	0.031	0.040
02	0.251	0.186	0.187	0.292	0.183	0.158	0.095	0.115	0.124
03	0.154	0.247	0.216	0.300	0.238	0.293	0.163	0.078	0.092
04	0.214	0.302	0.127	0.150	0.250	0.125	0.067	0.128	0.144
05	0.214	0.346	0.158	0.163	0.205	0.217	0.106	0.111	0.127
06	0.204	0.164	0.162	0.194	0.211	0.274	0.073	0.129	0.134
07	0.131	0.237	0.163	0.154	0.163	0.227	0.044	0.135	0.141
08	0.141	0.212	0.169	0.165	0.204	0.165	0.065	0.111	0.114
09	0.146	0.197	0.180	0.155	0.181	0.176	0.081	0.091	0.093
10	0.146	0.194	0.166	0.136	0.225	0.178	0.043	0.131	0.134
11	0.093	0.285	0.177	0.151	0.145	0.146	0.039	0.127	0.140
12	0.116	0.198	0.172	0.162	0.115	0.191	0.040	0.119	0.123
13	0.101	0.162	0.255	0.203	0.197	0.224	0.125	0.065	0.082
14	0.143	0.156	0.257	0.267	0.247	0.237	0.179	0.038	0.063
15	0.165	0.204	0.151	0.218	0.206	0.162	0.188	-0.004	0.026

4. The Effect of Only Two Sets of Samples Taking the Common Test Items on FCIP Linking Results

The common item parameter estimates were calibrated using a sample of 2000 subjects randomly chosen from the population. The final linking results were obtained by holding the common item parameter estimates fixed while estimating the rest of the item parameter estimates. The relationship between the results of FCIP based on the 2000-sample and the population are as highly correlated as 0.996, 0.999 and 0.987, respectively, for the item discrimination, difficulty and guessing parameter.

The subsequent EQUBANK run transformed the metric derived from the 2000-sample into that from the population by using a α value of 1.0001 and a β value of 0.0071. These transformation coefficients indicate that both scales are almost identical. Residuals were then computed by taking the differences between the linking parameters from the population and those from the 2000-sample. The average residuals are -0.002, 0.002 and -0.001, and their standard deviations are 0.024, 0.029 and 0.010 for the item discrimination, item difficulty and guessing parameter, respectively. These results indicate that the present IRT linking design can still be applied to equate just two tests.

B. Simulation Study

1. Comparisons of the Ability Parameters Equated by FCIP and CCM

The BIAS and the RMSE for each ability estimate were separately computed according to Equations III-1 and III-2. Their descriptive statistics computed across the 3000 examinees under FCIP and CCM are presented in Table IV-3. The average BIAS's of the 3000 examinees were found to be -0.069 and -0.045 for FCIP and CCM, respectively. Their corresponding standard deviations were 0.148 and 0.152, respectively. Likewise, the average RMSE were 0.304 and 0.302 for FCIP and CCM, respectively. Their corresponding standard deviations were 0.114 and 0.117, respectively. The plots of BIAS against the true ability parameters are presented in Figures 9 and 10 for FCIP and CCM, in that order. Similarly, the plots for RMSE are presented in Figures 11 and 12. The results indicate that both equating methods can produce very precise ability estimates for about 99 percent of the examinees. The BIAS statistics for most examinees are close to zero, except that those with extremely low ability ($\theta < -2.0$) were overestimated and those with extremely high ability ($\theta > 2.0$) were underestimated. FCIP, as compared to CCM, does not reduce BIAS when equating ability estimates. However, the ability estimates equated by FCIP appears to be slightly more stable than those by CCM. The reason why extremely high ability students have larger RMSE values is because the tests used for simulation are relatively easier.

Table IV-3
Descriptive Statistics of BIAS and RMSE for the Ability Parameter Estimated by FCIP and CCM. (N =3000)

	M	SD	Skewness	Kurtosis	Minimum	Maximum
Bias						
FCIP	-0.069	0.148	-4.100	41.586	-1.873	1.436
CCM	-0.045	0.152	-4.479	44.458	-1.933	1.447
RMSE						
FCIP	0.304	0.114	5.940	54.528	0.134	1.876
CCM	0.302	0.117	6.182	58.349	0.140	1.938

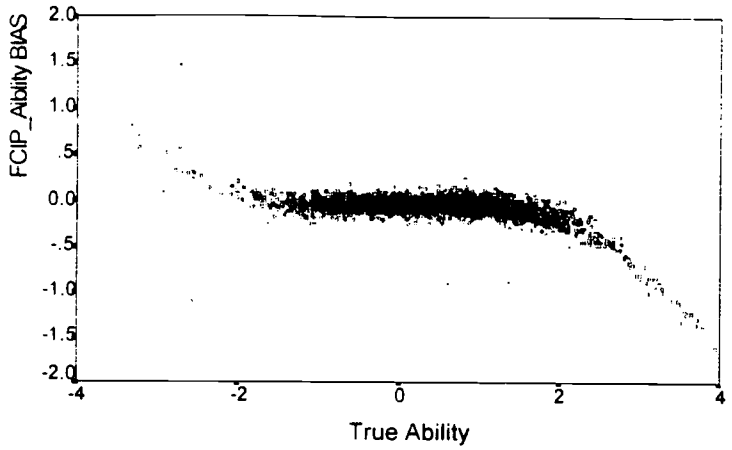


Figure IV-9: The Plot of the Ability Bias Statistics Produced by FCIP Versus the True Ability Parameters

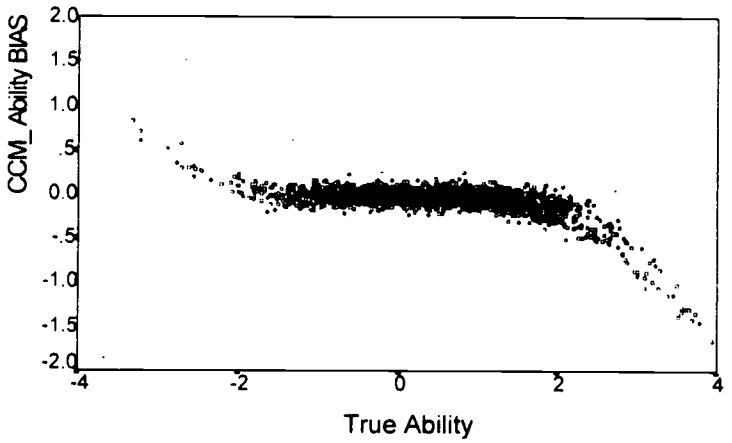


Figure IV-10: The Plot of the Ability Bias Statistics Produced by CCM Versus the True Ability Parameters

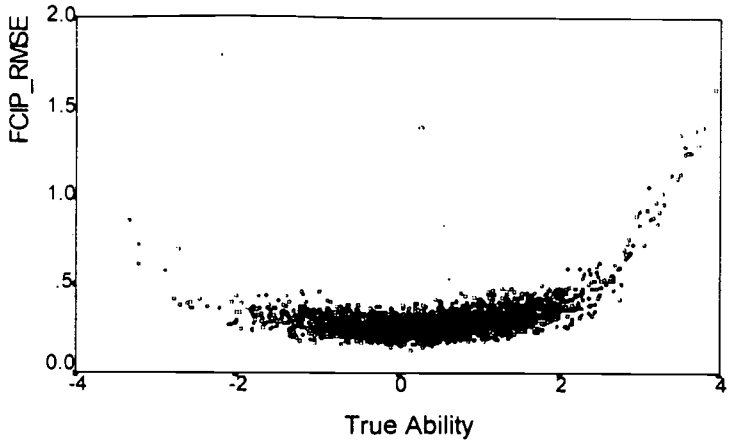


Figure IV-11: The Plot of the RMSE Statistics Produced by FCIP Versus the True Ability Parameters

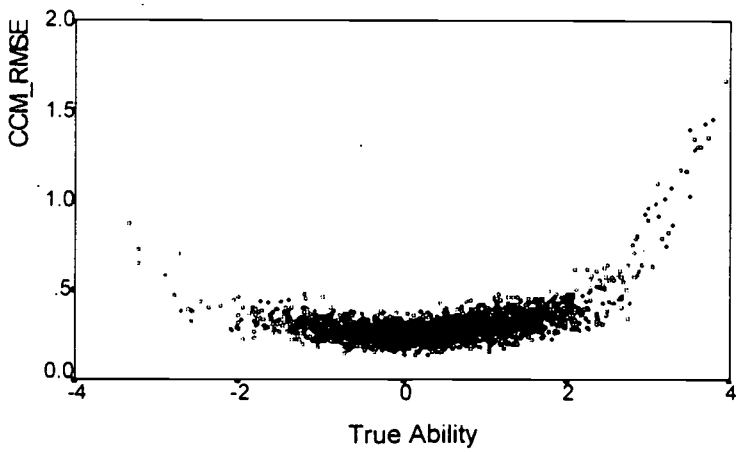


Figure IV-12; The Plot of the RMSE Statistics Produced by CCM Versus the True Ability Parameters

2. Comparisons of the Item Parameters Linked by FCIP and CCM

Using the Equations III-1 and III-2, the BIAS and RMSE indices for each item were separately computed. The descriptive statistics of these two indices computed across the 128 items under FCIP and CCM are presented in Table IV-4. In terms of the BIAS and RMSE indices, both equating methods can produce very stable and precise parameter estimates.

For the item discrimination estimates, the average BIAS's of the 128 items were found to be 0.019 and 0.015 for FCIP and CCM, respectively. Their corresponding standard deviations were 0.61 and 0.059, respectively. As for the average RMSE, they amounted to 0.132 and 0.132 for FCIP and CCM, respectively. The standard deviations were 0.052 and 0.050, respectively.

As regards the item difficulty estimates, the average BIAS's were -0.017 and - 0.014, and the standard deviations were 0.090 and 0.099 for FCIP and CCM, respectively. Likewise, the average RMSE's were 0.156 and 0.163 and their standard deviations were 0.076 and 0.085 for the two methods.

For the guessing parameters, the average BIAS's were 0.044 and 0.029 for the the two equating methods. The corresponding standard deviations were 0.039 and 0.045. The average RMSE's were 0.063 and 0.058 while their standard deviations were 0.039 and 0.029 for FCIP and CCM, respectively. The plots in Figures IV-13 to IV-15 below present the relationship between the BIAS statistics from FCIP versus the corresponding estimates of the item discrimination, difficulty and guessing parameters, in that order. The plots in Figures IV-16 to IV-18 are similar plots involving the RMSE statistics instead. The plots for CCM are similar to those under FCIP and are not presented here.

Table IV-4
Descriptive Statistics of the BIAS and RMSE Indices for the Various Item Parameter Estimated by FCIP and CCM. (Number of Items =128)

	a		b		c	
Bias	Mean	SD	Mean	SD	Mean	SD
Bias						
FCIP	0.019	0.061	-0.017	0.090	0.044	0.039
CCM	0.015	0.059	-0.014	0.099	0.029	0.045
RMSE						
FCIP	0.132	0.052	0.156	0.076	0.063	0.029
CCM	0.132	0.050	0.163	0.085	0.058	0.030

This simulation study was conducted under a combination of conditions:

1. The test data were generated from tests with easier items. This condition does not cater for precise estimation of the difficulty and guessing parameters.

2 .The metric conversion between the true and the estimated values constitutes a vertical linking situation. With these features in mind, further comparison can be made in the following.

The major difference between FCIP and other linking methods such as CCM is that FCIP includes the information about guessing during metric conversion between two tests. One hypothesis entertained in this study is that FCIP can produce more stable parameter estimates than CCM in case of an easier test because FCIP takes advantages of more information, including the guessing parameter and a larger sample size. Results from this study indicates that FCIP did produce a slightly more stable parameter estimates than CCM at the price of slight increases in the bias terms. One reasonable interpretation of these results is that the mean difficulty of the common items is less than 0.6 standard deviation from the mean difficulty for the rest of the test items. This reflects a vertical linking situation under FCIP, which usually produces more equating errors than horizontal linking. The metric transformation by CCM, on the other hand, is that of a horizontal linking.

So far, these two methods perform very well. Under other conditions (e.g. less relatively easy items), these two equating methods may perform even better.

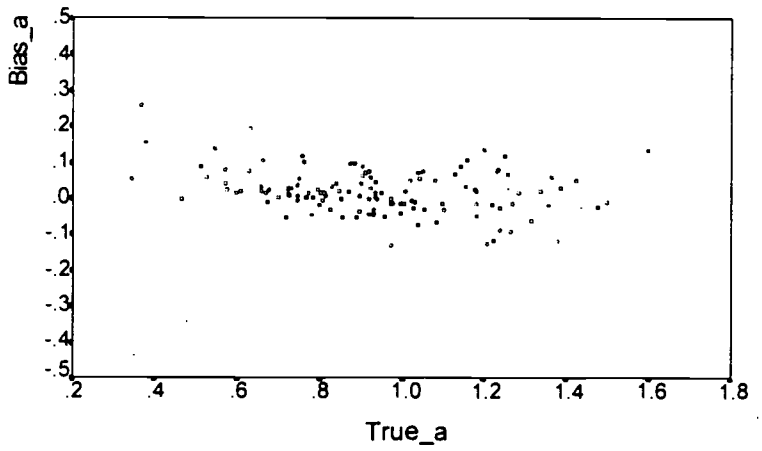


Figure IV-13. The Plot of the BIAS a Parameters Versus the True a Parameters

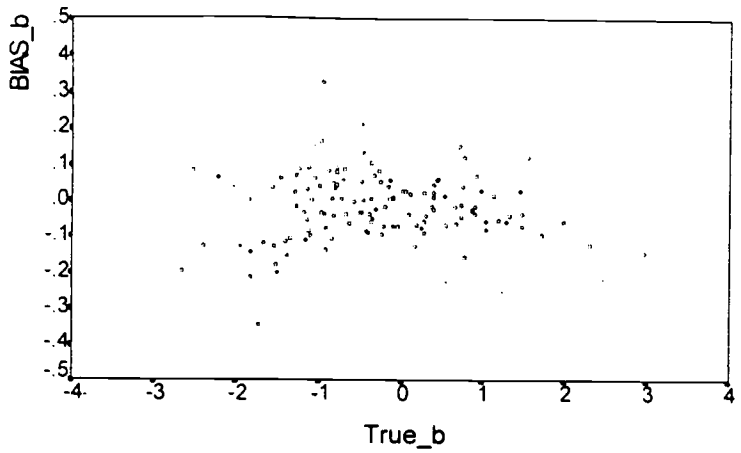


Figure IV-14. The Plot of the BIAS b Parameters Versus the True b Parameters

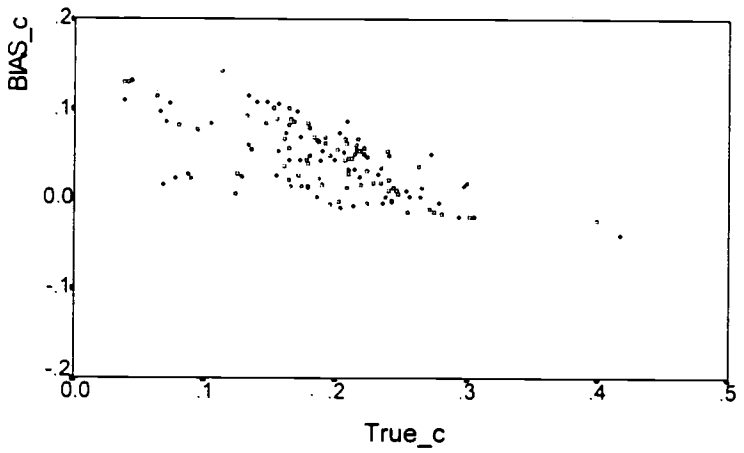


Figure IV-15. The Plot of BIAS c Parameters Versus the True c Parameters

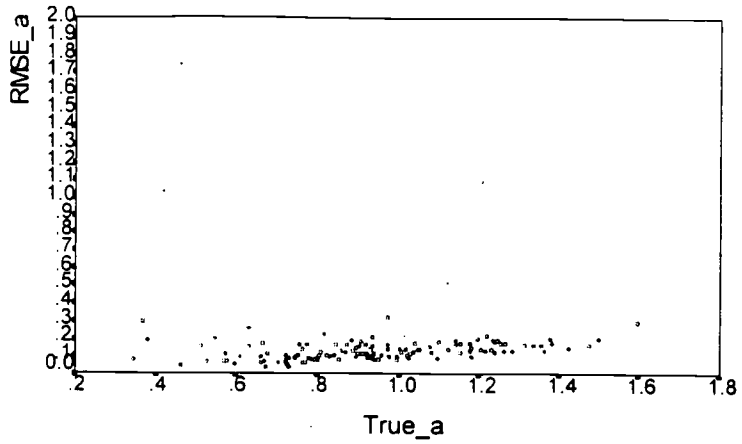


Figure IV-16: The Plot of the RMSE a Parameters Versus the True a Parameters

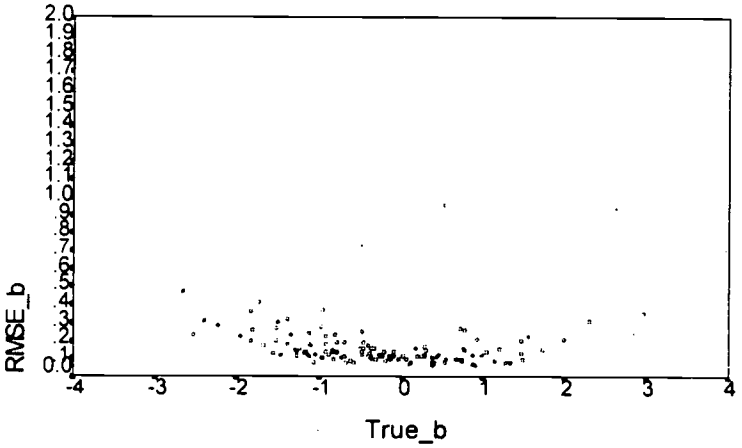


Figure IV-17: The Plot of the RMSE b Parameters Versus the True b Parameters

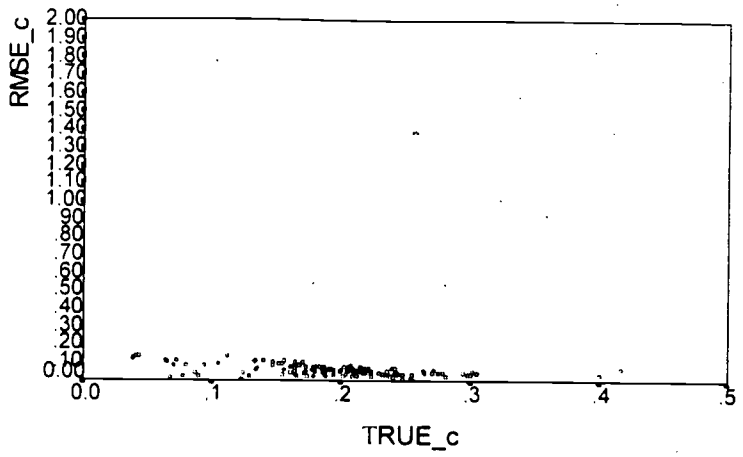


Figure IV-18: The Plot of the RMSE c Parameters Versus the True c Parameters

V. Applications and Conclusions

This study has explored the relative merits of a potentially useful IRT linking design. The procedures for this linking design are as follows:

1. Creation of Pilot Test Forms: Each form contains the same set of common test items.
2. Temporary Calibration of Each Pilot Test Form: Each pilot test form is calibrated and questionable items are removed.
3. Estimating the Common Item Parameters: The common test items are calibrated from the accumulated samples across all test forms.
4. Linking Items: Each pilot test is then calibrated again by fixing the common item parameter estimates to those resulting from the third step above, while estimating the rest of the items in one BILOG run.

After completion of the above procedures, the item parameter estimates calibrated from different tests and the ability estimates calibrated from different test forms are transformed onto the same scale.

With reference to the above discussion, this linking design has many nice practical features, including saving test developers time and energy in developing the pilot test forms, thereby allowing them to concentrate their energy in constructing good quality common test items across all the pilot tests. Based on the empirical portion of this study, ability estimates calibrated from this linking design are very consistent, except for students with extreme (especially low) ability under the CCM equating method. Item parameter estimates calibrated from this linking design are also very consistent, except for guessing parameters under the CCM equating method.

Based on the results from the simulation portion of the study, this linking result can produce very precise and stable parameter estimates. This simulation was conducted under the circumstances of larger standard errors for the item difficulty and guessing parameters of many

easier items, and that the mean difficulty of the common items is less than 0.5 standard deviation from the mean difficulty of the rest of the test items. This situation is reflective of the vertical linking framework of the FCIP linking method. It appears that the FCIP linking method is very robust and may potentially be used more extensively in building item banks. Whether the FCIP is robust in the other conditions (e.g. under heterogeneous groups that differ not only in terms of location, but also in variability) needs to be further examined in future studies.

References

Ackerman, T. A. (1987, April). The robustness of LOGIST and BILOG IRT estimation program to violations of local independence. Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24,1987).

Algina, J. (1986, April). A comparison of item parameter estimates and ability parameter estimates obtained by different methods implemented by BILOG. Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).

Baker, F. B. (1990). Some observations on the metric of PC-BILOG results. Applied Psychological Measurement, 14, 139-150.

Baker, F. B. (1992). Item Response Theory: Parameter Estimation Techniques. New York: Marcel Dekker, Inc.

Baker, F. B.(1993). EQUATE 2.0: A computer program for the test characteristic curve method of IRT equating. Applied Psychological Measurement, 17, 20.

Baker, F. B. (1996). An investigation of the sampling distributions of equating coefficients. Applied Psychological Measurement, 20, 45-57.

Baker, F. B. & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. Journal of Educational Measurement, 28, 147-162.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrica, 46, 443-458.

Brennan, R. (1987). Introduction to problems, perspectives, and practical issues in equating. Applied Psychological Measurement, 11, pp. 221-224.

Cook, L. L. & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), Applications of item response theory(pp. 175-195). Vancouver: Educational Research Institute of British Columbia.

Cook, L. L. & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optional circumstances. Applied Psychological Measurement, 11, pp. 225-244.

De Ayala, R. J., Schafer, W. & Sava-Bolesta, M. (1995). An investigation of the standard errors of expected a posteriori ability estimates. British Journal of Mathematical and Statistical Psychology, 47, 385-405.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. Applied Psychological Measurement, 9, pp. 413-415.

Elementary Mathematics Outcomes (1992). Prince George's County Public Schools, MD.

Hambleton, R. K. & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), New horizons in testing(pp.31-49). New York: Academic Press.

Hambleton, R. K. & Murray, L. N. (1983). Some goodness of fit investigations for item response models[1,2,3]. In R. K. Hambleton (Ed.), Applications of item response theory(pp.71-94). Vancouver: Educational Research Institute of British Columbia.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer Nijhoff Publishing.

Hambleton, R. K. & Swaminathan, H. Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Harwell, M. R. & Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. Applied Psychological Measurement, 15, 375-389.

Harwell, M. R. & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. Applied Psychological Measurement, 15, 279-291.

Kim, S. -H., & Cohen, A. S. (in press). A comparison of linking and concurrent calibration under item response theory. Applied Psychological Measurement.

Kim, S. & Cohen, A. S. (1997, March). A comparison of linking and concurrent calibration under the graded response model. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Kolen, M. J. & Brennan, R. L. (1995). Test equating: Methods and practices. New York: Springer-Verlag.

Li, Y. H. (1997). EQUBANK: A computer program to estimate the equating coefficients of the test characteristic curve method, Unpublished manuscript.

Li, Y. H. (1996). GEN3PL01: A Computer program to generate the one-PL, two- PL and three-PL IRT response patterns, Unpublished manuscript.

Linn, R., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

Loyd, B. H. & Hoover, H. D. (1980). Vertical equating using the Rasch Model. Journal of Educational Measurement, 17, 179-193.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. New Jersey: Lawrence Erlbaum Associates, Inc.

McKinley, R. L. & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. Applied Psychological Measurement, 9, 49-57.

Mislevy, R. J. & Bock R. D. (1982). Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program. In: Item Response Theory and Computerized Adaptive Testing Conference Proceedings (Wayzata, MN, July 27-30, 1982).

Mislevy, R. J. & Bock, R. D. (1984). Item operating characteristics of the Armed Services Vocational Aptitude Battery (ASVAB), Form 8a. Psychological Sciences Div., Office of Naval Research, Washington, D.C.

Mislevy, R. J. & Bock, R. D. (1990). BILOG-3 (2nd ed.): Item analysis and test scoring with binary logistic models. Mooresvilk: Scientific Software.

Mislevy, R. J. & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. Applied Psychological Measurement, 13, 57-75.

Nakamura, S. (1996). Numerical analysis and graphic visualization with MATLAB. Upper Saddle River, NJ: Prentice-Hall, Inc.

Petersen, N. S. Kolen, M. J. & Hoover H. D. (1989). Scaling, norming, and equating. In R. L. Linn (3rd Ed.), Educational Measurement, (pp. 221-262). New York: Macmillan.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. Applied Psychological Measurement, 14, 127-137.

Rogers H. J. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. Applied Psychological Measurement, 11, 47-57.

Seong, T. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. Applied Psychological Measurement, 14, 1990.

Skaggs, G. & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. Applied Psychological Measurement, 12, pp. 69-82.

Swaminathan, H. & Gifford, J. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), New horizons in testing(pp.13-30). New York: Academic Press.

Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. Psychometrika, 3, 461-475.

Stocking, M. L. Lord, F. M. (1983). Developing a common metric in items response theory. Applied Psychological Measurement, 7, 201-210.

Tam, H. P. & Li, Y. H. (1997, March). Is the use of the difference likelihood ratio chi-square statistics for comparing nested IRT models justifiable? Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

The MathWorks, Inc. (1995). MATLAB: The ultimate computing environment for technical education. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Thissen, D. & Wainer, H. (1982). Some standard errors in item response theory. Psychometrika, 47, 397-412.

Vale, C. D. (1986). Linking item parameters onto a common scale. Applied Psychological Measurement, 10, 333-344.

van der Linden, W. J. & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden and R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 1-28). New York: Springer-Verlag

Yen, W. M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5 245-262.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. Psychometrika, 52, 275-291.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Equating Multiple Tests via an IRT Linking Design: Utilizing a Single Set of Anchor Items with Fixed Common Item Parameters during the Calibration	
Author(s): Li, Yuan H., Griffith, William D. & Tam, Hak P.	
Corporate Source:	Publication Date: 3/17/98

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Yuan H. Li

William D. Griffith & Hak Tam

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

X

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2A

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2B

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Yuanhwan Li	Printed Name/Position/Title: Yuan H. LI Statistician
Organization/Address: Prince George's County Public Schools, Upper Marboro, MD. 20772 Room 205	Telephone: 301-952-6764 FAX: 301-952-6228
	E-Mail Address: yuanhwan@wam.umd.edu Date: 3/17/98

(over)